DOCUMENT RESUME

ED 222 579 TM 820 756

AUTHOR Fennessey, James; Salganik, Laura Hersh

TITLE Credible Comparison of Instructional Program Impact: The RAGS Procedure. Report No. 328.

INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for Social Organization of Schools.

SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Aug 82
GRANT NIE-G-0080-8
NOTE 26p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; Elementary Secondary Education; *Models; *Pretests Posttests; Program Effectiveness; Reliability; Research Problems; *Research Tools; Statistical Analysis; Test Interpretation
IDENTIFIERS *Rescaled and Adjusted Gains within Stratum

ABSTRACT

An explicit model identifying 10 relevant components of achievement gain scores has been developed. Based on that model, all students under consideration are stratified according to individual observed pretest score, and achievement gains are measured relative to the average and range of gains among students in the same prescore stratum. The resulting index, RAGS, is based on the Rescaled and Adjusted Gains within Stratum. Stratification by prescore controls well for the the biases identified in the decomposition of gain scores, and so allows the fairest practical comparison of program impacts. The RAGS reports also provide other data that allow educational managers to compare detailed impact patterns. By viewing the RAGS indices as useful approximations, and by institutionalizing a systematic procedure for critiquing and refining the index construction process, educators have available one major component of an overall program assessment system that is informative, feasible, and self-improving. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for Social Organization of Schools

Report No. 328
August 1982

CREDIBLE COMPARISON OF INSTRUCTIONAL PROGRAM IMPACT: THE RAGS PROCEDURE

James Fennessey and Laura Hersh Salganik

The Johns Hopkins University

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Hollifield

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

STAFF

Edward L. McDill, Co-Director
James M. McPartland, Co-Director

Karl L. Alexander
Charles H. Beady
Henry J. Becker
Jomills H. Braddock, II
Ruth H. Carter
Martha A. Cook
Robert L. Crain
Doris R. Entwisle
Joyce L. Epstein
Gail M. Fennessey
James J. Fennessey
Homer D. C. Garcia
Denise C. Gottfredson
Gary D. Gottfredson
Linda S. Gottfredson
Stephen Hansell
Edward J. Harsch
John H. Hollifield
Barbara J. Hucksoll
Nancy L. Karweit
Hazel G. Kennedy
Marshall B. Leavey
Nancy A. Madden
David J. Mangefrida
Julia B. McClellan
Anne McLaren
Phillip R. Morgan
Robert G. Newby
Deborah K. Ogawa
James M. Richards, Jr.
Donald C. Rickert, Jr.
Laura Hersh Salganik
Robert E. Slavin
Gail E. Thomas
William T. Trent
Carol A. Weinreich

CREDIBLE COMPARISON OF INSTRUCTIONAL PROGRAM IMPACT: THE RAGS PROCEDURE

Grant No. NIE-G-0080-8

James Fennessey and Laura Hersh Salganik

Report No. 328

August 1982

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States National Institute of Education, Department of Education. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the Institute should be inferred.

Center for Social Organization of Schools
The Johns Hopkins University
3505 North Charles Street
Baltimore, MD 21218

Printed and assembled by the Centers for the Handicapped
Silver Spring, MD

Introductory Statement

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The Center works through five programs to achieve its objectives. The Studies in School Desegregation program applies the basic theories of social organization of schools to study the internal conditions of desegregated schools, the feasibility of alternative desegregation policies, and the interrelations of school desegregation with other equity issues such as housing and job desegregation. The School Organization program is currently concerned with authority-control structures, task structures, reward systems, and peer group processes in schools. It has produced a large-scale study of the effects of open schools, has developed Student Team Learning Instructional processes for teaching various subjects in elementary and secondary schools, and has produced a computerized system for school-wide attendance monitoring. The School Process and Career Development program is studying transitions from high school to post secondary institutions and the role of schooling in the development of career plans and the actualization of labor market outcomes. The Studies in Delinquency and School Environments program is examining the interaction of school environments, school experiences, and individual characteristics in relation to in-school and later-life delinquency.

The Center also supports a Fellowships in Education Research program that provides opportunities for talented young researchers to conduct and publish significant research, and to encourage the participation of women and minorities in research on education.

This report presents a procedure for using gain scores to compare the impact of different educational programs.

Abstract

An explicit model identifying ten relevant components of achievement gain scores has been developed. Based on that model, all students under consideration are stratified according to individual observed pretest score, and achievement gains are measured relative to the average and range of gains among students in the same prescore stratum. The resulting index, RAGS, is based on the Rescaled and Adjusted Gains within Stratum.

Stratification by prescore controls well for the biases identified in the decomposition of gain scores, and so allows the fairest practical comparison of program impacts. The RAGS reports also provide other data that allow educational managers to compare detailed impact patterns.

By viewing the RAGS indices as useful approximations, and by institutionalizing a systematic procedure for critiquing and refining the index construction process, educators have available one major component of an overall program assessment system that is informative, feasible, and self-improving.

OBJECTIVES AND APPROACH

In recent years, achievement test scores have been used increasingly by education professionals and interested citizens as evidence of the quality of education. How well a school is doing or how well an instructional program works often is judged largely by whether its students' scores are "going up" or "going down," or whether they are "high" or "low" relative to national norms. But such score patterns alone are hardly credible as indicators of instructional program impact. Students entering a program with many relevant skills may learn little from the instruction and still finish with relatively high scores, while less well-prepared students may learn a great deal and still finish with relatively low scores. Intuitively, gain scores -- the difference between a student's score at the end of instruction and her/his score at the beginning -- are the obvious candidate for describing program impact. Measurement of program impact clearly must somehow incorporate measurement of knowledge gain. However, gain scores have been widely attacked by measurement technicians, and the conflict between intuitive and technical requirements has hampered productive use of test scores for program comparisons.

This paper presents a conceptualization of gain scores that we believe leads to a technically acceptable tool for comparing the impact of different instructional programs. We propose an index called RAGS, Rescaled and Adjusted Gains within Strata, along with a flexible but explicit procedure to compare program impacts. The entire RAGS framework has been designed to be adapted and refined by users in order to meet local information and management needs.

We stress, however, that no index based on achievement test results should be used alone in program evaluation or education management. A complete assessment procedure should include multiple sources of information, such as ratings by knowledgeable persons, classroom observations, or other test data <<footnote 1>>.

USING GAINS TO COMPARE PROGRAM IMPACT

This section places comparison of program impact in the context of a general model of gain scores. Two key concepts for the model are gain score and program. Gain score (GAIN) is the difference between a student's posttest score and pretest score. It can be positive or negative. Because our goal is to make comparisons, not to estimate the absolute level of impact, gain scores do not have to be based on the same test <<footnote 2>>. A program is the grouping of students that is to be compared, e.g. a district within a state, a region within a district, a school within a district, a grade within a school, or an instructional method. Comparisons can be subject-specific (math, reading) or can combine test results in different subjects.

A Gain Score Model

As illustrated in Equation 1, individual gain scores reflect several phenomena.

EQ (1) GAIN = IPI + RESP + REG + BND + DIS, where:

IPI: Instructional Program Impact is the component of gain attributable to the program unit to which the student belongs.

RESP: Responsiveness is the component of gain attributable to the

student's current average learning rate. It is the net consequence of learning ability, motivation, and relevant previous learning, and is in general positively associated with true gain scores. For example, students with high responsiveness will gain more, for any given program quality. This pattern is sometimes referred to as "fan-spread."

REG: Regression-to-the-mean is the component of gain attributable to the negative association between observed pretest scores and observed gain scores. For example, students with high pretest scores will tend to have high measurement errors -- i.e. inflated pretest scores -- and thus tend to have smaller observed gain scores, for any given program quality. Regression is a subtle and complex topic <<footnote 3>>, and will be addressed more fully in future work.

BND: Boundary Artifacts are the result of the tests being too easy or too hard for the students being tested. They deflate observed gains, and are often referred to as "floor and ceiling effects."

DIS: Disturbances are random fluctuations of observed scores that are logically and empirically unrelated to the other gain score components. They produce imprecision in estimates of other components (unreliability), and are also referred to as "noise."

Each student's gain can also be decomposed into a program average and a within-program deviation from the average. This is shown in Equation 2.

EQ (2) GAIN = AVG + DEV, where:

AVG: Program Average is the average gain score for students in the instructional program.

DEV: Deviation is the difference between the individual's score and

the program average.

Combining Equations 1 and 2 produces ten components that contribute to each student's gain score. This is shown in Table 1. The key item

Insert Table 1 About Here

of interest is IPIAVG, instructional impact at the program level. Note that in the model, program impact is completely identified as an average, and is separate from the IPIDEV component.

As Table 1 demonstrates, total average gain (AVG) reflects four phenomena substantively distinct from program impact. Furthermore, as discussed above, three of these components (REGAVG, RESPAVG, and BNDAVG) are not random, but instead are associated with average characteristics of individual students. For example, a low gain score can indicate low average responsiveness, accidentally high pretest scores, or floor/ceiling effects, in addition to low program impact.

Statistically, this means that in most situations, average gain (AVG) is a biased estimate of average program impact (IPIAVG); the expected value of AVG is not IPIAVG, but varies depending on the average pretest score of the students. Because different programs often have students with different pretest scores, the effects of the bias make comparisons among programs unfair. Thus, to estimate average impact (IPIAVG) fairly, we need to remove the three biasing components from AVG.

From Gain Scores to RAGS

10

At this point, it is important to recall the purpose of our endeavor. Any impact index should provide maximally fair and reasonably precise comparisons of instructional program impact. That is all we seek. This

means: (1) We need only to equalize bias before making comparisons, not to estimate the absolute level of bias. (2) We need only to equalize the net bias arising from all three sources, not to equalize the biasing components separately. (3) We need only to assume that groups are affected equally by bias, not that all individual scores are equally biased. This eases our task considerably. The problem then is to develop an index such that the aggregate net bias -- in terms of the model, RESPAVG + REGAVG + BNDAVG -- is essentially equal in all program groups. We do not need to control for differences in any of the deviation components -- RESPDEV, REGDEV, and BNDDEV -- because they do not contribute to bias in the average gain (AVG).

For practical purposes, the pretest score provides, in a single number, the best available information on all three of the biasing components just discussed. We can reasonably assume that any groups of students with similar pretest scores will have similar responsiveness and error biases, and thus will be subject to similar amounts of fan-spread, regression-to-the-mean, and floor or ceiling effects. In particular, instructional program subgroups which have similar pretest scores can be viewed as having similar aggregate net bias.

At this point, it is necessary to introduce the concept of reference population. This is defined as the total group of students enrolled in all the programs which will be compared. To develop the impact index, the reference population is divided into several strata according to students' pretest scores (e.g. deciles). Then, z-scores of gains are calculated for each student, using the mean and standard deviation of the raw gains from the prescore stratum to which that student belongs as

the adjusting and rescaling constants. Because we convert to z-scores within each prescore stratum separately, the effects of pretest score on gain are controlled. We call the resulting z-scores RAGS, because they are Rescaled and Adjusted Gain Scores within Stratum. After the RAGS scores are calculated for individual students, they can be aggregated and reported according to program units, such as grade within a school, total school, or subject. Given our assumptions, the RAGS averages for each program -- regardless of its prescore composition -- have an expected net bias of zero, and an expected value of IPIAVG in z-score form.

Two Illustrations

To illustrate how the adjusting and rescaling works, we will present two simple artificial examples, in which we use only two prescore control strata (low prescore and high prescore). In the first example, we compare the effectiveness of the math programs in three schools (A, B, and C). Table 2 shows that Program C has the highest raw average

Insert Table 2 About Here

gain (10.5) and program A the lowest (7.5). However, within each prescore stratum, the raw average gains are the same for all three programs. To control for the different prescore compositions, we therefore stratify each school by pretest score, and adjust each student's score by the average gain for students in the same prescore stratum. The adjusted averages show no difference in instructional impact among Programs A, B, and C. Each program now has an adjusted average gain of zero. The program total differences in raw average gain were due to their different mixes of students -- Programs B and C had a

higher percentage of students in the high prescore stratum, where gains were larger. With adjusted averages, Program A -- containing mostly students from a prescore stratum in which all gains are smaller -- does not suffer or benefit when compared to Programs B and C.

As long as the stratum standard deviations are the same (as they are in the example above), if two programs have equal impact, then the average of their adjusted scores will be the same. However, if the control strata have unequal standard deviations, and we neglect to rescale for them, comparisons among programs will be affected. Thus, the procedure of adjusting only for stratum means is incomplete. For example, suppose we are comparing two programs whose students have considerably different pretest score distributions, such as Programs A and B in Table 3. Suppose also that gains in the prescore strata to

Insert Table 3 About Here

which most students in Program A belong vary more widely than the gains in the strata containing most of Program B's students. In general, Program A will then have a larger adjusted average gain than Program B; this gives the false impression that Program A has more impact than Program B. The impression is false because the interpretation of any particular gain depends on the relevant <u>range</u> of gain scores. Therefore, we <u>rescale</u> the scores in each prescore stratum, dividing each gain by the stratum standard deviation.

In the example illustrated in Table 3, the standard deviations in the two strata are unequal (1.90 vs. 3.81). Based on comparisons of raw gains <u>within</u> prescore strata, Programs A and B have equal impact. The raw average gain for low prescore students is 8 and for high prescore

students is 16, regardless of whether they belong to A or B. However, looking at the Program Total column, Program A has a higher raw average gain (14) than program B (10) and a higher adjusted average gain (3.5 vs. 2.5). The first is caused by the different prescore compositions of the two programs. Most of Program A's students are in the high prescore group; most of Program B's are in the low prescore group. The second arises because most of Program A's students are in the prescore stratum with the wider range of scores. To adjust for this, we divide each student's adjusted average score by the standard deviation of her/his stratum. After this is done, the RAGs averages for Program A and Program B are the same (1.05), as our intuition requires. Similarly, Programs E and F have unequal raw and adjusted averages, but equal RAGS. Programs C and D have unequal raw averages but equal adjusted averages; rescaling doesn't affect their average gains. This is because the raw averages gains within each stratum are at the stratum means.

RAGS INDICES AND REPORTS

Describing the Rescaled and Adjusted Gains Distributions

The most straightforward index of average program impact is the average RAGS score of the students enrolled. The pattern or distribution of RAGS scores within each program, however, is also important. Information about distributions of gains is provided in two ways in RAGS summary reports.

First, the average RAGS are reported separately for subgroups of students within an instructional program — for example, those with low, middle, and high pretest scores. This information is useful because

reporting only the RAGS average for a program can obscure meaningful differences among programs. For example, a program whose RAGS average is relatively high might have produced excellent results with its brightest students, but mediocre results with the slower students. Another program with the same RAGS average might have had equal impacts on all its students. These differences in impact patterns help educators identify how the program affects different groups of students. The RAGS summary report shows them quite clearly, a feature that is much appreciated by administrators.

These reporting subgroups are independent of the control strata used to calculate the RAGS. Thus, presentation of results separately for different subgroups of students is a readily generalizable feature. If an administrator wished, and the necessary information were available, RAGS results could be presented separately for boys and girls, for blacks and whites, or for students with low and high absence rates, etc.

Second, the distribution of RAGS within a program is further described by showing the standard deviation of the RAGS for all students in each program, and for each subgroup of students for which average RAGS are reported. In addition, any RAGS distributions showing unusual shape (as measured by skewness and kurtosis) are flagged for further attention.

Showing the Imprecision in the RAGS Indices

Because RAGS reports on programs are based on aggregated scores, they are more precise as estimates of impact than the corresponding individual scores. However, some random error certainly remains, especially if the individual scores used to calculate the index have low reliability, and if the number of students in the reporting group is small.

We are developing an indicator of the degree of imprecision in a RAGS program average, based on the pooled within-strata variances for that program. This imprecision estimate includes random measurement error deriving from the individual pretests and postests, error arising from imperfect control of program group differences, and error due to additional program-level sources of extraneous disturbances (e.g. fire drills just before testing) <<footnote 4>>.

Implementing the RAGS Procedure

Our approach assigns several specific tasks to the administrators who draw information from it. In order to compute RAGS scores, they must specify (1) the reference population, (2) the prescore control strata, and (3) the instructional program groupings and reporting subgroups within programs. A mathematics specialist with district-wide responsibilities probably would use a reference population consisting of all students in the district who are taking mathematics. A regional supervisor might also use the district-wide reference population, but then ask for RAGS summary reports only on the schools in her/his region.

Designating the prescore control strata involves selecting the prescore cutpoints which will be used to define "equivalent" prescores. This requires a preliminary technical examination of gains distributions for separate prescore values <<footnote 5>>. Practitioners familiar with testing can assist others in this task. Eventually, these cutpoints will be suggested by the RAGS computer program.

Once RAGS scores have been computed for individual students, practitioners must specify the instructional programs for which RAGS distributions will be reported. Program groupings could be defined by any characteristic believed to influence instructional impact, e.g. the method of instruction, the instructional materials, or the management strategy used, as well as by grade within a school. Also, the users may designate subgroups of students within programs whose RAGS distributions will be reported separately. For example, our work to date has provided separate results for students with prescores in the low, the middle, and the high third of the reference population.

Implementing the RAGS Computations

A system of FORTRAN computer modules for carrying out the RAGS procedure and preparing the summary reports is now undergoing final testing. To help users designate prescore control strata cutpoints -- i.e., selecting the ranges of pretest scores to be regarded as equivalent -- one module generates a table showing the distributions of gains in the reference population for each possible prestest score. Other modules allow the user to apply alternate definitions of reference populations, prescore control strata, instructional programs, and reporting subgroups of students within programs. A report generator

module allows flexibility in the report layout and detail provided. The FORTRAN system also includes a file creation and management module for defining and maintaining a master longitudinal file of test information. These program modules can be adapted for most computers, including some of the more powerful microprocessors.

FURTHER CONSIDERATIONS

Limitations and Potential Refinements

Any index based on gain scores as a measure of individual response to an instructional program has two conceptually distinct aspects, which usually are not made explicit. First, the gain score allows the analyst to think in terms of change in the student's knowledge during the instructional program. This is the basis of its wide intuitive appeal. Second, the prescore is a reasonable choice as a control variable for removing major biasing factors, especially the differences between individual students in their responsiveness to instruction, (i.e., their expected rate of gain during instruction).

These controls, however, are not perfect. First, partitioning the reference population by the observed prescore is not completely equivalent to partitioning on the true pretest score. If the pretest reliability is very low, there is little real partitioning, and nearly as much true prescore variation remains within each stratum as was originally found in the entire reference population. Conversely, if the prescore is highly reliable, then the partitioning, though not perfect, is quite good.

Second, prescore is an imperfect proxy for responsiveness because a student's true prescore reflects not only the student's responsiveness, but also that student's cumulated actual response to the instruction she/he has received prior to the pretest. It is an imputation of current learning rate based on observed past learning rates. Clearly, however, all students have not received an identical mixture of instruction in their prior education. Some have been exposed to material that was too difficult and/or too fast-paced. Others have not been sufficiently challenged. In each case, the likely bias in prescore as an indicator of responsiveness is to portray some students as being less responsive than they are.

Finally, to the extent that regression-to-the-mean and test boundary effects are not distributed evenly across programs within each prescore stratum, there is residual bias remaining in the RAGS averages. These, however, are technical problems affecting <u>all</u> procedures which use observed variables as controls, including linear regression methods that control on observed (rather than true) measures of SES, IQ, and/or achievement tests.

Fundamentally, the dilemma is inescapable. Instructional program impact, responsiveness, and measurement error jointly determine gain scores. To measure impact, responsiveness, <u>or</u> bias, one must be able to measure and thus control for the others. The sensible course is to accept the fact that any particular control variable will be imperfect, and then to do the best possible job using the control variables we do have. Despite the concerns just described (and we do not ignore or minimize them), we believe that the observed prescore is a good initial

choice as a control for responsiveness, regression-to-the-mean, and test boundary biases.

Ongoing refinement is an integral part of the RAGS framework. As assessment is repeated periodically, users in each local setting will gradually agree about conceptual elements and will incorporate additional information into strata definitions to improve the credibility of the RAGS reports. Those elements could well include: a student's scores on tests taken in earlier years, the student's attendance patterns, and perhaps cognitive ability measures, socioeconomic background measures, previous school grades or teachers' ratings, etc. Such multiple quantitative sources of evidence about responsiveness can be introduced gradually but systematically; discrepancies between one source and another can be brought to light and discussed. Strata boundaries might even be made overlapping (as a way to reduce the problems of small sample sizes). Still other definitions of the reference population are also possible. For example, the reference population eventually might include the performance of students in prior years as well as the current group. Finally, RAGS can be computed using several alternative control strata and reference populations with the same test scores. These results can be compared in an informal, pragmatic sensitivity analysis. The flexibility of the RAGS framework facilitates such local refinement and gradual evolution of the procedures.

It is also easy to broaden the RAGS assessment to include criterion-referenced features by using different rescaling and adjusting constants. Instead of using <u>observed</u> average and standard deviation

within each stratum as the adjusting and rescaling constants, the constants could be an average and standard deviation based on goals stated by parents and/or practitioners. More generally, it is possible to combine externally designated stratum means and sigmas with those observed in the reference population.

Using RAGS Wisely

RAGS users should understand both the inherent limitations of any quantitative index of this kind, and the asumptions implicit in using such an index. First, RAGS are not a definitive final measurement, but instead only one component of a system for comparing instructional impact. That system uses multiple sources of evidence. Second, RAGS averages do not compare the average gains of students in one stratum with those in another; all stratum RAGS averages are zero. Differences between strata in average level of gain thus must be considered in a different analysis. Third, RAGS distributions per se provide no information about absolute levels of achievement. For this reason, the RAGS Summary report also presents averages on prescores, postscores, and gains.

Last, to avoid potential mistrust, abuse, or loss of focus, the RAGS procedure needs periodic systematic review and refinement by competent and concerned users. The reviewers must be knowledgable, respected, and able to devote adequate time and resources to this task.

SUMMARY

21

RAGS reports show several indices of instructional program impact based on the distribution of rescaled and adjusted achievement gains

within each program. These indices are created by stratifying a reference population by the observed prescores, and then measuring all gains relative to the average gain of students in the same prescore stratum. Thus, the RAGS approach is a form of norm-referenced scoring, with norms based on a reference population designated by local educators.

Although stratification by prescore is not a perfect control for bias, it produces essentially fair comparisons of program impact, despite differences in starting characteristics of students and measurement problems associated with gain scores. RAGS reports allow educators to compare fairly and in some detail the instructional impact of different programs.

By regarding the RAGS indices as useful approximations and by systematically critiquing and refining the index construction procedure, users can employ the RAGS framework as a major component of an assessment system that is informative, feasible, and self-improving.

FOOTNOTES

1. This procedure was developed as part of the Instructional Program Analysis Project (IPAP), funded by NIE. The project investigates a broad systems-based strategy for assessing the effectiveness of instructional programs. To avoid the overreliance on one source of evidence -- test scores or personal impressions -- the IPAP project seeks to systematically integrate multiple sources of information about program impact, permitting a "triangulation" of evidence. When several sources of information indicate similar levels of program impact, they corroborate each other, and thus increase the credibility of the assessment. Conversely, if the sources are discrepant, this identifies areas where additional field studies, or management effort to clarify the meaning of the evidence, would be valuable. The IPAP project also includes structured feedback from a panel of local educators about the assessment procedure itself. This helps guide evolutionary refinement and the development of more useful, efficient procedures.

2. In fact, the pretest and postest need only be congeneric. That is, they need only measure the same concept, but may have differing origins, scales, and error variances. This means that various score formats, as well as various test forms, can be used. Raw score is usually simplest and most precise, but the procedure also can be applied to other score formats such as grade-equivalents or NCE scores.

3. Users of test scores often understand regression (or regression-to-the-mean) as a result of measurement error in the tests. However, this is too simple. In fact, any observed regression to the mean is a composite of regression in the measurement error component, and of regression in the true score component. Classification of a set of observed scores into low, middle, and high thus involves classification by both the error components and the true score components. Therefore, the interpretation of regression patterns becomes complex, especially if subgroups of students have different true score means on any component.

4. Additional sources of error include unbalanced N's across strata and programs.

5. Users can also exclude students for whom they judge floor and/or ceiling effects to be unacceptably large.

Table 1

Decomposition of Individual Gain Scores

	AVG program average	DEV within-program individual deviations
IPI (program impact)	IPIAVG	IPIDEV
RESP (responsiveness)	RESPAVG	RESPDEV
REG (regression)	REGAVG	REGDEV
BND (boundary effects)	BNDAVG	BNDDEV
DIS (disturbances)	DISAVG	DISDEV

Table 2

Raw and Adjusted Gains by Instructional Program and Stratum

	Pretest Score Stratum LOW	Pretest Score Stratum HIGH	Program Total
Program A	RA = 6 AA = 0 SD = 1 N = 15	RA = 12 AA = 0 SD = 1 N = 5	RA = 7.5 AA = 0 SD = 2.8 N = 20
Program B	RA = 6 AA = 0 SD = 1 N = 10	RA = 12 AA = 0 SD = 1 N = 10	RA = 9.0 AA = 0 SD = 3.2 N = 20
Program C	RA = 6 AA = 0 SD = 1 N = 5	RA = 12 AA = 0 SD = 1 N = 15	RA =10.5 AA = 0 SD = 2.8 N = 20
Reference Population	AV = 6 SD = 1 N = 30	AV = 12 SD = 1 N = 30	

RA=raw average SD=standard deviation N=number of students
AA=adjusted average

Table 3

Raw Gains, Adjusted Gains, and RAGS, by Program and Stratum

	Prescore Strata LOW	Prescore Strata HIGH	Program Total
Program A	RA = 8 SD = 1 N = 5 AA = +2	RA = 16 SD = 2 N = 15 AA = +4	RA = 14.00 RAGAVG = +1.05 SD = 3.97 N = 20 AA = 3.5
Program B	RA = 8 SD = 1 N = 15 AA = +2	RA = 16 SD = 2 N = 5 AA = +4	RA = 10.00 RAGAVG = +1.05 SD = 3.77 N = 20 AA = 2.5
Program C	RA = 6 SD = 1 N = 5 AA = 0	RA = 12 SD = 2 N = 15 AA = 0	RA = 10.50 RAGAVG = 0 SD = 3.20 N = 20 AA = 0
Program D	RA = 6 SD = 1 N = 15 AA = 0	RA = 12 SD = 2 N = 5 AA = 0	RA = 7.50 RAGAVG = 0 SD = 2.95 N = 20 AA = 0
Program E	RA = 4 SD = 1 N = 5 AA = -2	RA = 8 SD = 2 N = 15 AA = -4	RA = 7.00 RAGAVG = -1.05 SD = 2.51 N = 20 AA = -3.5
Program F	RA = 4 SD = 1 N = 15 AA = -2	RA = 8 SD = 2 N = 5 AA = -4	RA = 5.00 RAGAVG = -1.05 SD = 2.18 N = 20 AA = -2.5
Reference Population	AV = 6 SD = 1.90 N = 60	AV = 12 SD = 3.81 N = 60	

RA=raw average SD=standard deviation N=number of students
AA=adjusted average RAGAVG=RAGS average